Robust Estimation of Inverse Probability Weights for Marginal Structural Models*

Kosuke Imai† Marc Ratkovic‡

Forthcoming in
Journal of the American Statistical Association

Abstract

Marginal structural models (MSMs) are becoming increasingly popular as a tool to make causal inference from longitudinal data. Unlike standard regression models, MSMs can adjust for time-dependent observed confounders while avoiding the bias due to the adjustment for covariates affected by the treatment. Despite their theoretical appeal, a main practical difficulty of MSMs is the required estimation of inverse probability weights. Previous studies have found that MSMs can be highly sensitive to misspecification of treatment assignment model even when the number of time periods is moderate. To address this problem, we generalize the Covariate Balancing Propensity Score (CBPS) methodology of Imai and Ratkovic (2014) to longitudinal analysis settings. The CBPS estimates the inverse probability weights such that the resulting covariate balance is improved. Unlike the standard approach, the proposed methodology incorporates all covariate balancing conditions across multiple time periods. Since the number of these conditions grows exponentially as the number of time period increases, we also propose a low-rank approximation in order to ease the computational burden. Our simulation and empirical studies suggest that the CBPS significantly improves the empirical performance of MSMs by making the treatment assignment model more robust to misspecification. Open-source software is available for implementing the proposed methods.

Key Words: causal inference, covariate balancing propensity score, inverse propensity score weighting, observational studies, sequential ignorability, time-dependent treatments

---

*The proposed methodology can be implemented via open-source software `CBPS` (Fong *et al.*, 2014), which is freely available as an R package at the Comprehensive R Archive Network (CRAN http://cran.r-project.org/package=CBPS). We thank seminar participants at Rutgers University (Statistics and Biostatistics), the University of Michigan (Economics), the University of St. Gallen (Economics), the University of Wisconsin (Biostatistics), and the Atlantic Causal Inference Conference (Harvard University) for helpful suggestions.

†Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: http://imai.princeton.edu

‡Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 608–658–9665, Email: ratkovic@princeton.edu, URL: http://www.princeton.edu/∼ratkovic

1 Introduction

Since its introduction by Robins (1999), marginal structural models (MSMs) have quickly gained popularity among applied researchers in biomedical and other fields as a tool for making causal inference from longitudinal data in observational studies. The paper that popularized MSMs in the field of epidemiology has more than 1,000 Google Citations as of March 2014 (Robins *et al.*, 2000) and the method has been introduced to other disciplines (e.g., Blackwell, 2013). As explained by Robins *et al.* (2000), when estimating the causal effects of time-varying treatments, standard regression models fail to appropriately adjust for time-dependent observed confounders that are affected by previous treatments. In contrast, MSMs allow one to estimate the causal effects of different treatment sequences while avoiding this post-treatment bias.

Despite their theoretical appeal, a main practical difficulty of MSMs is the required estimation of inverse probability weights. Using simulation and empirical studies, a number of previous studies have found that MSMs can be highly sensitive to misspecification of treatment assignment model even when the number of time periods is moderate (e.g., Cole and Hernán, 2008; Howe *et al.*, 2011; Kang and Schafer, 2007; Lefebvre *et al.*, 2008; Mortimer *et al.*, 2005). The effect of misspecification can propagate across time periods because the inverse probability weights used for MSMs are typically based on the product of propensity score estimated separately at each time period.

To address this problem, we introduce the Covariate Balancing Propensity Score (CBPS) methodology as an alternative estimation method for inverse probability weights of MSMs. The CBPS was first introduced by Imai and Ratkovic (2014) to improve the estimation of propensity score in the cross section settings. In this paper, we generalize the CBPS methodology to longitudinal data. In the cross-

sectional case, inverse probability weights reduce confounding bias through balancing pre-treatment covariates between treated and untreated observations. We extend this logic to the longitudinal setting. We show that, at every time period, MSM weights must balance across all potential *future* treatment assignments, conditional on the *past* treatment assignment. Therefore, unlike the standard approach, the proposed methodology incorporates all covariate balancing conditions when estimating inverse probability weights. We then use these balance conditions as estimating equations. The resultant weights are robust in the sense that they improve covariate balance even when the treatment assignment model is misspecified.

After briefly reviewing MSMs and their assumptions (Section 2), we describe the proposed CBPS methodology (Section 3). We then conduct simulation studies to show that the CBPS can dramatically improve the empirical performance of MSMs when the treatment assignment model is misspecified (Section 4). In addition, we present an empirical application to show that the CBPS achieves a greater degree of covariate balance than the standard approach and yields substantively different results (Section 5). The final section gives concluding remarks and discusses future research agenda.

2 A Review of Marginal Structural Models

In this section, we briefly review the marginal structural models (MSMs) of Robins (1999). See Robins *et al.* (2000) and Blackwell (2013) for more detailed introduction of MSMs. Suppose that we have a simple random sample of size n from a population. For each unit $i = 1, 2, \ldots, n$, repeated measurements are taken at each of J time periods. Specifically, at each time period $j = 1, 2, \ldots, J$, we observe the time-dependent treatment variable T_{ij} as well as the time-dependent confounders X_{ij} that are possibly affected by previous treatments. We assume that X_{ij} is realized before the treatment

at time j and therefore is not affected by T_{ij}. We further assume that the treatment variable is binary where $T_{ij} = 1$ ($T_{ij} = 0$) implies unit i receives (does not receive) the treatment at time j. Next, for each unit, we denote the treatment and covariate history up to time j by $\overline{T}_{ij} = \{T_{i1}, T_{i2}, \ldots, T_{ij}\}$ and $\overline{X}_{ij} = \{X_{i1}, X_{i2}, \ldots, X_{ij}\}$, respectively. We also denote the set of possible treatment and covariate values at time j as $\overline{\mathcal{T}}_j$ and $\overline{\mathcal{X}}_j$. Finally, we observe the outcome variable Y_i for unit i at the end of the study, i.e., time J, after the treatment for the same time period, i.e., T_{iJ}, is administered.

The potential outcome framework of causal inference was originally developed by Neyman (1923) and Rubin (1973) in the cross-section setting, but Robins (1986) generalized it to the longitudinal analysis. Under this framework, we use $Y_i(\bar{t}_J)$ to represent the potential value of the eventual outcome variable for unit i measured at time J under the entire treatment history $\overline{T}_{iJ} = \bar{t}_J$ where $\bar{t}_J \in \overline{\mathcal{T}}_J$. Thus, the observed outcome is given by $Y_i = Y_i(\overline{T}_J)$. Similarly, $X_{ij}(\bar{t}_{j-1})$ denotes the potential values of covariates for unit i at each time period j under the treatment history up to time $j-1$, i.e., $\overline{T}_{i,j-1} = \bar{t}_{j-1}$. Therefore, the observed values of covariates can be written as $X_{ij} = X_{ij}(\overline{T}_{i,j-1})$ for unit i at time j. This setup relies upon the consistency assumption that the potential values of outcome and covariates for each unit are only functions of its own treatment history up to that point in time. The assumption excludes the possible interference between units (but not between time periods), implying that the potential values of outcome and covariates are not influenced by the treatment history of other units.

MSMs are based on the assumption of sequential ignorability, which states that the treatment assignment of unit i at time j is exogenous given the treatment and covariate history of the same unit up to that point in time. In other words, MSMs assume no unmeasured confounding at each time period. This sequential ignorability

assumption can be formally written as,

$$Y_i(\bar{t}_J) \perp\!\!\!\perp T_{ij} \quad | \quad \overline{T}_{i,j-1} = \bar{t}_{j-1}, \overline{X}_{ij} = \bar{x}_j \tag{1}$$

at any time period j for a given treatment history $\bar{t}_J = \{\bar{t}_{j-1}, t_j, \ldots, t_J\} \in \overline{\mathcal{T}}_J$ and covariate history $\bar{x}_j \in \overline{\mathcal{X}}_j$. We also assume that the conditional probability of treatment assignment is bounded away from zero and one at each time period. That is,

$$0 \; < \; \Pr(T_{ij} = 1 \mid \overline{T}_{i,j-1} = \bar{t}_{j-1}, \overline{X}_{ij} = \bar{x}_j) \; < \; 1 \tag{2}$$

at any time period j for a given treatment history $\bar{t}_{j-1} \in \overline{\mathcal{T}}_{j-1}$ and covariate history $\bar{x}_j \in \overline{\mathcal{X}}_j$.

Under these assumptions, Robins (1999) showed that the inverse-probability-of-treatment weighting can be used to consistently estimate the marginal mean of any potential outcome, i.e., $\mathbb{E}\{Y_i(\bar{t}_J)\}$ for any treatment sequence $\bar{t}_J \in \mathcal{T}_J$. For the reason that will become clear later, we first define the potential value of this weight for unit i under treatment history $\bar{t}_J$ as,

$$w_i(\bar{t}_J, \overline{X}_{iJ}(\bar{t}_{J-1})) \;\; = \;\; \frac{1}{P(\overline{T}_{iJ} = \bar{t}_J \mid \overline{X}_{iJ}(\bar{t}_{J-1}))} \;\; = \;\; \prod_{j=1}^{J} \frac{1}{P(T_{ij} = t_{ij} \mid \overline{T}_{i,j-1} = \bar{t}_{j-1}, \overline{X}_{ij}(\bar{t}_{j-1}))} \tag{3}$$

This weight is typically small and therefore the estimates become highly variable. Therefore, researchers commonly follow the suggestion given in the literature and use the stabilized weights of the form, $w_i^*(\bar{t}_J, \overline{X}_{iJ}(\bar{t}_{J-1})) = P(\overline{T}_{iJ} = \bar{t}_J)/P(\overline{T}_{iJ} = \bar{t}_J \mid \overline{X}_{iJ}(\bar{t}_{J-1}))$, when fitting the outcome model. We denote the observed values of these weights as $w_i = w_i(\overline{T}_{iJ}, \overline{X}_{iJ})$ and $w_i^* = w_i^*(\overline{T}_{iJ}, \overline{X}_{iJ})$.

In an observational study, these weights are unknown and must be estimated. Typically, a parametric model is used to estimate the conditional probability of treatment assignments given the set of covariates,

$$w_i^{-1} \;\; = \;\; \pi_\beta(\overline{T}_{iJ}, \overline{X}_{iJ}) \tag{4}$$

where β is a finite dimensional vector of unknown parameters. A common choice of parametric model is the logistic regression independently applied to each time period,

$$\pi_\beta(\overline{T}_{iJ}, \overline{X}_{iJ}) \;=\; \prod_{j=1}^{J} \text{expit}\{(2T_{ij}-1)\beta_j^\top \overline{V}_{ij}^*\} \tag{5}$$

where $\overline{V}_{ij}^* = [\overline{T}_{i,j-1}\ \overline{X}_{ij}]$, $\text{expit}(z) = \{1+\exp(-z)\}^{-1}$, and β_j is a vector of unknown coefficients. The numerator of the stabilized weight is typically estimated using the sample proportion for each treatment sequence.

Once the (stabilized) weights are estimated, the conditional expectation of outcome is modeled as a function of treatment history alone without covariates, i.e., $\mathbb{E}(Y_i \mid \overline{T}_{iJ})$. For example, researchers may regress the outcome on the treatment indicators from all periods. Robins (1999) has shown that this weighting approach yields a consistent estimate of the mean potential outcome, i.e., $\mathbb{E}\{Y_i(\bar{t}_J)\}$ thereby allowing researchers to compute the average outcome under any sequence of treatment assignments over time.

3 The Proposed Methodology

In this section, we propose an alternative estimation procedure for the inverse-probability-treatment weight for MSMs. Specifically, unlike the standard approach, we estimate the weight such that time-dependent covariates are balanced across all appropriate sub-populations. The proposed methodology generalizes the the covariate balancing propensity score (CBPS) of Imai and Ratkovic (2014) to the longitudinal data settings. The key idea of the CBPS is to estimate the propensity score such that the resulting covariate balance is improved. Therefore, the CBPS is robust in the sense that even under a misspecified treatment assignment model the covariate balancing conditions, which are used as estimating equations, are improved. In addition, since the proposed methodology focuses on the estimation of the MSM weights, it can be combined with other approaches to achieve the double-robustness property (Yu and

van der Laan, 2006) (see also Graham *et al.* (2012) who develop a doubly-robust estimator in the cross-section setting). We begin by reviewing this methodology and then show how to extend the CBPS to the causal analysis with panel data.

3.1 The Single Time Period Case: A Review

We first review the CBPS in the cross-section setting. Imai and Ratkovic (2014) propose to estimate the propensity score model such that the following covariate balance condition is satisfied,

$$\mathbb{E}\left\{\frac{T_i X_i}{\pi_\beta(1, X_i)} - \frac{(1-T_i)X_i}{\pi_\beta(0, X_i)}\right\} \;=\; 0. \tag{6}$$

Imai and Ratkovic suggest that these moment conditions can be used to estimate the propensity score model either via generalized method of moments or empirical likelihood. Their simulation and empirical studies find that the CBPS significantly improves the performance of standard propensity score estimation. Several other methods have also been developed to improve covariate balance (e.g., Hainmueller, 2012; Graham *et al.*, 2012), but to the best of our knowledge, none has dealt with longitudinal data settings, to which we now turn.

3.2 The Two Time Period Case

To convey the intuition for the proposed methodology, we first present the CBPS for the case of two time periods before discussing the general case of more than two time periods. Suppose that for each unit i, we observe the outcome variable Y_i measured at the end of study, the binary treatment variable T_{ij}, and a vector of confounders X_{ij} for each time period $j = 1, 2$. Further assume that we are interested in using MSMs to estimate the marginal mean of potential outcome measured at the end of the second period, $\mathbb{E}\{Y_i(\bar{t}_2)\}$, where $\bar{t}_2$ can take any of the four possible values, i.e., $\bar{t}_2 \in \mathcal{T}_2 = \{(0,0), (0,1), (1,0), (1,1)\}$.

Covariate Balancing Conditions. We derive the moment conditions based on the covariate balancing property of the weight for MSMs. To do this, we first express these moment conditions as functions of the (potential) weight defined in equation (3). Specifically, at the first time period, across all four possible treatment histories, the weight should balance the mean of the baseline covariate, X_{i1}. Formally, for all $\bar{t}_2 = (t_1, t_2) \in \{(0,0), (0,1), (1,0), (1,1)\}$, we have

$$\mathbb{E}(X_{i1}) \;=\; \mathbb{E}\left[\mathbf{1}\{T_{i1} = t_1, T_{i2} = t_2\} w_i(\bar{t}_2, \overline{X}_{i2}(t_1)) X_{i1}\right]. \tag{7}$$

This gives the total of three moment conditions because the above equality holds across four different treatment histories and one such equality is redundant.

While there exist numerous equivalent ways to represent these three moment conditions, we choose the following orthogonal representation, which can be written in a compact notation using the observed weight instead of its potential values,

$$\mathbb{E}\{(-1)^{T_{i1}} w_i X_{i1}\} \;=\; 0 \tag{8}$$

$$\mathbb{E}\{(-1)^{T_{i2}} w_i X_{i1}\} \;=\; 0 \tag{9}$$

$$\mathbb{E}\{(-1)^{T_{i1}+T_{i2}} w_i X_{i1}\} \;=\; 0 \tag{10}$$

This orthogonal representation of covariate balancing conditions is summarized in the first three rows of Table 1. In the table, if we treat + and − as +1 and −1, row vectors for each time period are orthogonal to each other.

The covariate balancing conditions at the second time period are similar to those at time 1, except that the covariates measured at time 2 are possibly functions of the treatment at time 1, i.e., $X_{i2} = X_{i2}(T_{i1})$. This means that the covariate balancing conditions will be conditional on the observed treatment value at time 1. Using the potential outcomes notation, for all $\bar{t}_2 = \{t_1, t_2\}$, we can write these covariate balancing conditions as follows,

$$\mathbb{E}\{X_{i2}(t_1)\} \;=\; \mathbb{E}\left[\mathbf{1}\{T_{i1} = t_1, T_{i2} = t_2\} w_i(\bar{t}_2, \overline{X}_{i2}(t_1)) X_{i2}(t_1)\right] \tag{11}$$

	Treatment history: (t_1, t_2)				
Time period	(0,0)	(0,1)	(1,0)	(1,1)	Moment condition
time 1	+	+	−	−	$\mathbb{E}\left\{(-1)^{T_{i1}} w_i X_{i1}\right\} = 0$
	+	−	+	−	$\mathbb{E}\left\{(-1)^{T_{i2}} w_i X_{i1}\right\} = 0$
	+	−	−	+	$\mathbb{E}\left\{(-1)^{T_{i1}+T_{i2}} w_i X_{i1}\right\} = 0$
time 2	+	−	+	−	$\mathbb{E}\left\{(-1)^{T_{i2}} w_i X_{i2}\right\} = 0$
	+	−	−	+	$\mathbb{E}\left\{(-1)^{T_{i1}+T_{i2}} w_i X_{i2}\right\} = 0$

Table 1: Orthogonal Representation of Covariate Balancing Moment Conditions in the Two Time Period Case. The first and second time periods have three and two moment conditions, respectively. There are four distinct values of treatment history with t_j representing the value of the treatment variable at time j. The symbols, "+" and "−", in these four treatment history columns show whether the weighted average of covariates among units with a certain treatment history is added or subtracted when formulating the moment condition. Within each time period, row vectors of +'s and −'s for the treatment history combinations are orthogonal to one another. The last column represents the corresponding moment condition.

Because $X_{i2}(t_1)$ is observed only when $T_{i1} = t_1$, the above covariate balancing equation implies that X_{i2} should be balanced across treatment values at time 2 conditional on the treatment value realized at time 1.

Similar to the baseline covariate case, we use the orthogonal representation, which in this case yields the following two moment conditions,

$$\mathbb{E}\left\{(-1)^{T_{i2}} w_i X_{i2}\right\} = 0 \tag{12}$$

$$\mathbb{E}\left\{(-1)^{T_{i1}+T_{i2}} w_i X_{i2}\right\} = 0 \tag{13}$$

The bottom two rows of Table 1 summarize this result. While at time 1 both T_{i1} and T_{i2} are varied to generate 3 moment conditions, only T_{i2} is varied at time 2. By now, readers may realize the benefit of our orthogonal representation: as shown in Section 3.3, its compact notation allows one to easily extend the proposed methodology to the general case of more than two time periods.

Estimation. Since the number of moment conditions exceeds the number of parameters to be estimated, we use the generalized method of moments (GMM; Hansen, 1982) estimation to combine the covariate balancing conditions derived above. Our

optimal GMM estimator is given by,

$$\hat{\beta} \;=\; \underset{\beta\in\Theta}{\operatorname{argmin}}\ \operatorname{vec}(\mathbf{G})^{\top}\mathbf{W}^{-1}\operatorname{vec}(\mathbf{G}) \tag{14}$$

where the sample moment conditions are given by,

$$\mathbf{G} \;=\; \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} (-1)^{T_{i1}}w_iX_{i1} & (-1)^{T_{i2}}w_iX_{i1} & (-1)^{T_{i1}+T_{i2}}w_iX_{i1} \\ 0 & (-1)^{T_{i2}}w_iX_{i2} & (-1)^{T_{i1}+T_{i2}}w_iX_{i2} \end{bmatrix}, \tag{15}$$

and their covariance $\mathbf{W}$ is given by,

$$\mathbf{W} \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\begin{bmatrix} 1 & (-1)^{T_{i1}+T_{i2}} & (-1)^{T_{i2}} \\ (-1)^{T_{i1}+T_{i2}} & 1 & (-1)^{T_{i1}} \\ (-1)^{T_{i2}} & (-1)^{T_{i1}} & 1 \end{bmatrix}\otimes w_i^2\begin{bmatrix} X_{i1}X_{i1}^{\top} & X_{i1}X_{i2}^{\top} \\ X_{i2}X_{i1}^{\top} & X_{i2}X_{i2}^{\top} \end{bmatrix}\;\middle|\; X_{i1}, X_{i2}\right\} \tag{16}$$

The expectation in equation (16) can be calculated analytically, for example, for the logistic regression case (Imai and Ratkovic, 2014).

3.3 The General Longitudinal Case

We now extend the above formulation to the general case with more than two time periods, i.e., $j = 1, 2, \ldots, J$. We first generalize the covariate balancing conditions derived above and then propose the optimal GMM estimator. We also consider its low-rank approximation to address a computational challenge when the number of time periods is large.

Covariate Balancing Conditions. We characterize the covariate balancing conditions in the general case with an arbitrary number of time periods $J \geq 2$. Recall that in the two time period case, the weight for MSMs balances the covariates at the first time period across all potential values of the entire treatment vector. At the second time period, however, the weight only balances covariates across the treatment values at that time period among the units who receive the same treatment value in

the first time period. In general, the weight balances covariates at a given time period across all potential combinations of the current and future treatment conditions given the past treatment sequence.

Formally, for a given time period j and fixed past treatment sequence up to that point $\bar{t}_{j-1}$, we can write the covariate balancing conditions across all treatment sequences of the current and future time periods $\underline{t}_j = \{t_j, t_{j+1}, \ldots, t_J\}$ as,

$$\mathbb{E}\{X_{ij}(\bar{t}_{j-1})\} \;=\; \mathbb{E}[\mathbf{1}\{\overline{T}_{j-1} = \bar{t}_{j-1}, \underline{T}_{ij} = \underline{t}_j\} w_i(\bar{t}_J, \overline{X}_{iJ}(\bar{t}_{J-1})) X_{ij}(\bar{t}_{j-1})] \quad (17)$$

where $\underline{T}_{ij} = \{T_{ij}, T_{i,j+1}, \ldots, T_{iJ}\}$ represents a vector of observed current and future treatment conditions.

In the two time period case, the balance conditions are characterized in terms of the sums and differences of $w_i X_{ij}$ across all groups defined by a distinct value of the entire treatment sequence. We generalize that formulation here. Specifically, for each time period, we use the orthogonal representation of the covariate balancing conditions given in equation (17) by aliasing the past treatment effects on the covariates at time j. Since there exist a total of 2^{J-j+1} potential current and future treatments, equation (17) implies $2^{J-j+1} - 1$ orthogonal constraints given a particular history of treatment up to time $j-1$, i.e., $\bar{t}_{j-1}$. There are a total of 2^{j-1} possible treatment histories and hence all together we have $(2^J - 2^{j-1})$ covariate balancing conditions for each time period j.

To formalize this idea, we utilize the theoretical framework developed for analyzing and designing randomized experiments based on the 2^J full factorial design (see e.g., Box *et al.*, 2005). In Table 2, we present a running example of the case with $J = 3$ where the first three columns present the design matrix in Yates order with $+$'s and $-$'s indicating the presence and absence of the treatment at each time period, respectively. It is well recognized that the full 2^J factorial design can be represented by Hadamard matrix of order 2^J. Recall that Hadamard matrix of order n, denoted

Design matrix			Treatment history: (t_1, t_2, t_3)								Time periods		
T_{i1}	T_{i2}	T_{i3}	(0,0,0) h_0	(1,0,0) h_1	(0,1,0) h_2	(1,1,0) h_{12}	(0,0,1) h_{13}	(1,0,1) h_3	(0,1,1) h_{23}	(1,1,1) h_{123}	1	2	3
−	−	−	+	+	+	+	+	+	+	+	✗	✗	✗
+	−	−	+	−	+	−	+	−	+	−	✓	✗	✗
−	+	−	+	+	−	−	+	+	−	−	✓	✓	✗
+	+	−	+	−	−	+	+	−	−	+	✓	✓	✗
−	−	+	+	+	+	+	−	−	−	−	✓	✓	✓
+	−	+	+	−	+	−	−	+	−	+	✓	✓	✓
−	+	+	+	+	−	−	−	−	+	+	✓	✓	✓
+	+	+	+	−	−	+	−	+	+	−	✓	✓	✓

Table 2: Orthogonal Representation of Covariate Balancing Moment Conditions in the Three Time Period Case Using the 2^3 Factorial Experiment Framework. The first three columns show the design matrix of the factorial experiment in Yates order where the symbol "+" ("−") represents the presence (absence) of each treatment factor. The next eight columns show the Hadamard matrix of this factorial experiment based on this design matrix that corresponds to the eight distinct values of treatment history with t_j representing the value of the treatment variable at time j. The symbols, "+" and "−", in these eight treatment history columns also indicate the orthogonal representation of covariate balancing moment conditions. Finally, the symbol ✓(✗) in the last three columns indicates that the corresponding covariate balancing moment condition is (not) binding for each time period.

by $\mathbf{H}_n$, is an $n \times n$ matrix of $+1$'s and -1's whose rows are orthogonal to one another, implying that $\mathbf{H}_n^\top \mathbf{H}_n = n\mathbf{I}_n$.

To construct a Hadamard matrix that corresponds to the full 2^J factorial design, let $\mathbf{D}$ be the $2^J \times J$ "negative" design matrix of $+1$'s and -1's sorted in Yates order,

$$\mathbf{D} \quad = \quad [d_0, d_1, d_2, d_{12}, d_3, d_{13}, d_{23}, d_{123}, d_4, d_{14}, \ldots]^\top \tag{18}$$

where d_0 is a J dimensional column vector of 1's and d_j is a column vector of length J where the elements of set j indicate the indexes of the elements of the vector with -1 and the other elements of the vector are 1's. For example, when $J = 3$, we have $d_{12} = (-1, -1, 1)^\top$. Thus, +'s and −'s in Table 2 correspond to -1's and $+1$'s in $\mathbf{D}$, respectively. Let c_j be the jth column of $\mathbf{D}$ so that $\mathbf{D} = [c_1, c_2, \ldots, c_J]$.

Further, denote the common component of d_j and c_k by d_{jk}. For a subset t of $\mathbb{N}_J = \{1, \ldots, J\}$, let the Hadamard product, denoted by h_t, of columns c_k with $k \in t$ be a 2^J dimensional column vector with its jth element being $\prod_{k \in t} d_{jk}$. Then, the

Hadamard matrix of order 2^J can be constructed by collecting in Yates order all the Hadamard products of the columns of $\mathbf{D}$. The result is given by the following $2^J \times 2^J$ matrix,

$$\mathbf{H}_{2^J} = [h_0, h_1, h_2, h_{12}, h_3, h_{13}, h_{23}, h_{123}, h_4, h_{14}, \ldots] \tag{19}$$

where h_0 is a column vector of $+1$'s. This matrix in the case of $J = 3$ is given in the middle columns of Table 2.

The Hadamard matrix representation allows us to enumerate all the covariate balancing moment conditions in a systematic way regardless of the number of time periods. Moreover, the successive multiplication procedure used for the construction of this Hadamard matrix directly justifies the notation used in equations (12) and (13). In fact, it has long been known that this Hadamard matrix representation can be used to compute the mod 2 discrete Fourier transform (Good, 1958). For example, the second and sixth rows of Table 2 corresponds to the following covariate moment conditions,

$$\mathbb{E}\{(-1)^{T_{i1}} w_i X_{ij}\} = 0 \tag{20}$$

$$\mathbb{E}\{(-1)^{T_{i1}+T_{i3}} w_i X_{ij}\} = 0 \tag{21}$$

That is, one can use the design matrix to form the treatment variables that enter the exponent of -1 in the compact expression of the covariate balancing moment conditions. In sum, the 2^J factorial experiment framework allows us to directly generalize the orthogonal representation of the covariate moment conditions given in Section 3.2 to the general case with more than two time periods.

Therefore, this full 2^J factorial design framework clearly shows which covariate balancing moment conditions are binding at any given time period for the estimation of the weight for MSMs. As noted above, the stabilized weight balances covariates measured at time j across all possible current and future treatments but it does

not balance across past treatments. The covariate balancing moment conditions, which correspond to the effects of past treatments and their interactions, are not binding. These conditions can be easily identified by the design matrix. For example, in Table 2, we see that the second row, corresponding to the main effect of time 1 treatment, i.e., T_{i1}, is not binding for time 2 covariates X_{i2}. Similarly, for time 3 covariates, the moment conditions corresponding to the effects of T_{i1} and T_{i2} as well as their interaction are not binding. In general, for covariates measured at time j, the first 2^{j-1} rows of Hadamard matrix $\mathbf{H}_{2^J}$ can be ignored when constructing the covariate balancing moment conditions.

Estimation. As in the two time period case, we use the GMM to combine all the covariate balancing conditions. We begin by defining the following matrices for covariates,

$$\widetilde{\mathbf{X}} \;=\; \left[\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n\right]^\top \tag{22}$$

where $\widetilde{X}_i = [w_i X_{i1}, w_i X_{i2}, \ldots, w_i X_{iJ}]^\top$ is a $(K \times J)$ dimensional column vector of covariates for unit i. Next, we construct the $n \times (2^J - 1)$ model matrix based on the design matrix $\mathbf{D}$ arranged in Yates' order as

$$\mathbf{M} \;=\; [M_1,\ M_2, \ldots, M_n] \tag{23}$$

where $M_i = [m_{i0}, m_{i1}, m_{i2}, m_{i12}, m_{i3}, m_{i13}, m_{i23}, m_{i123}, m_{i4}, m_{i4}, \ldots]^\top$ is a $(2^J - 1)$ dimensional column vector with $m_{i0} = 1$ and $m_{it} = (-1)^{\sum_{k \in t} T_{ik}}$ for $t \in \{1, 2, 12, 3, 13, 23, 123, 4, 14, \ldots\}$. For example, m_{i23} equals $(-1)^{T_{i2}+T_{i3}}$ and m_{i123} equals $(-1)^{T_{i1}+T_{i2}+T_{i3}}$. In fact, the ith row of $\mathbf{M}$ is given by the row of the Hadamard matrix in Table 2 that corresponds to the treatment sequence of the ith observation.

Given this notation, our optimal GMM estimator is given by equation (14) with the following generalized definitions of the sample balancing condition and their con-

ditional covariance,

$$\mathbf{G} = \frac{1}{n}\sum_{i=1}^{n}\left(M_i^\top \otimes \widetilde{X}_i\right)\mathbf{R} \tag{24}$$

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(M_i M_i^\top \otimes \widetilde{X}_i\widetilde{X}_i^\top \mid X_i\right) \tag{25}$$

where $\mathbf{R}$ represents the "selection" matrix which identifies the binding covariate balancing conditions for each time period and "zeros" them out. This matrix is formally defined as,

$$\mathbf{R} = [\mathbf{R}_1\ \mathbf{R}_2\ \ldots \mathbf{R}_J] \quad \text{where} \quad \mathbf{R}_j = \begin{bmatrix} \mathbf{0}_{2^{j-1}\times 2^{j-1}} & \mathbf{0}_{2^{j-1}\times(2^J-2^{j-1})} \\ \mathbf{0}_{(2^J-2^{j-1})\times 2^{j-1}} & \mathbf{I}_{2^J-2^{j-1}} \end{bmatrix} \tag{26}$$

for each $j = 1, \ldots, J$. As mentioned earlier, the expectation in equation (25) can be evaluated analytically for the logistic regression case.

When the number of time periods is large, the inversion of $\mathbf{W}$ can be computationally expensive because its dimension, which is $\{(2^J-1)\times JK\}\times\{(2^J-1)\times JK\}$, exponentially increases as a function of J. Here, we derive a low-rank approximation to the full covariance matrix as a way to overcome this computational difficulty. To do this, we assume that the correlation across balance conditions is zero. Note that when this assumption does not hold the resulting GMM estimator is still consistent but no longer efficient. In our simulation and empirical studies (see Section 4 and 5), we find that the empirical performance is not greatly affected by this approximation especially in a large sample size.

Specifically, our low-rank approximation to the covariance matrix is given by,

$$\widetilde{\mathbf{W}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}\otimes\widetilde{X}_i\widetilde{X}_i^\top = \mathbf{I}\otimes\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}. \tag{27}$$

where the variances in this new matrix are identical to those in the original $\mathbf{W}$ matrix of equation (25) but certain covariances are zero. Then, our GMM estimator is given

by,

$$\hat{\beta} = \underset{\beta \in \Theta}{\operatorname{argmin}} \operatorname{vec}(\mathbf{G})^{\top} \{\mathbf{I} \otimes \widetilde{\mathbf{X}}^{\top} \widetilde{\mathbf{X}}\}^{-1} \operatorname{vec}(\mathbf{G}) \tag{28}$$

$$= \underset{\beta \in \Theta}{\operatorname{argmin}} \operatorname{trace}\{\mathbf{R}^{\top} \mathbf{M}^{\top} \widetilde{\mathbf{X}} (\widetilde{\mathbf{X}}^{\top} \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^{\top} \mathbf{M} \mathbf{R}\} \tag{29}$$

Thus, this approximation approach avoids the Kronecker product and the inversion of a large matrix.

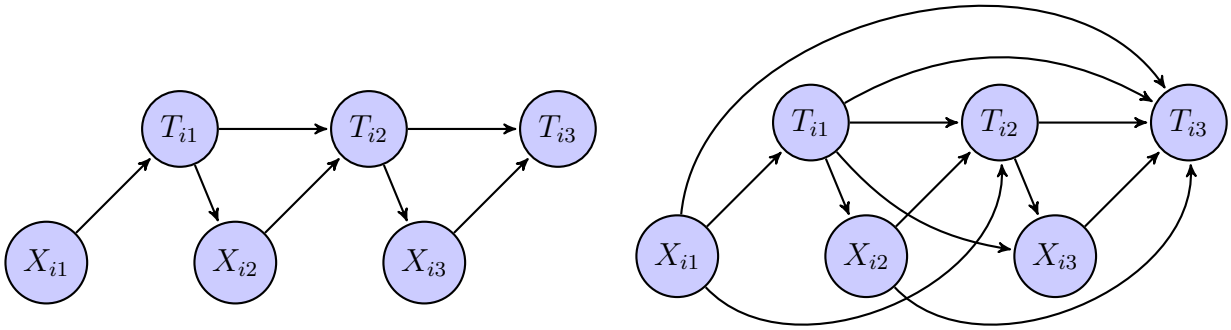
3.4 Extension to Multiple Binary Treatments

The method described above naturally extends to the setting where there exist multiple binary treatments. Indeed, dynamic treatment regimes considered in this paper is essentially a special case of J multiple binary treatments. The only difference is that for dynamic treatment regimes some of the covariate balancing conditions are not binding as indicated by zero elements of $\mathbf{G}$ matrix in equations (15) and (24). In contrast, for multiple binary treatments, all of these covariate balancing conditions are binding. However, aside from this difference, the estimation for the case of multiple binary treatments proceeds in an identical manner.

4 Simulation Studies

We conduct four sets of simulation studies in order to assess the empirical performance of the proposed CBPS estimation. First, we show that when the treatment assignment model is correctly specified, the proposed methodology does as well as the standard maximum likelihood estimation. Second, we also examine several scenarios where the treatment assignment model is misspecified in terms of either the lag structure or the functional form of the covariates (or both). We find that the CBPS significantly reduces the bias and mean squared error of the standard method in each of these model misspecification scenarios.

In all four simulation scenarios, we consider the case of three time periods, i.e.,



(a) Simulation 1

(b) Simulation 2

Figure 1: Treatment Variable Data Generating Process in Simulation Studies. In the first set of simulations summarized by the diagram of panel (a), a relatively simple treatment assignment model is used and we only misspecify the functional form while maintaining the correct lag structure. In the second set of simulations summarized by the diagram of panel (b), a more complex data generating process is used and we examine the impact of incorrectly specifying the lag structure. The results of these simulations are given in Figure 2 and 3, respectively

$J = 3$, and use four different sample sizes $n = 500, 1,000, 2,500$, and $5,000$. Across these four simulations, we vary both whether the lag structure and functional form for the treatment assignment model are properly modeled. Figure 1 summarizes the treatment variable data generating processes used in our simulations. In the first set of simulations summarized by the diagram of panel (a), a relatively simple treatment assignment model is used and we only misspecify the functional form while maintaining the correct lag structure. The treatment-generating process in this setup is a function of exogenous covariates and the immediately previous observed treatment level.

In practice, however, both the treatment variables and the covariates may be affected by more than the immediately previous time period. In the second set of simulations, summarized by the diagram of panel (b), a more complex data generating process is used and we examine the impact of incorrectly specifying the lag structure. The treatment-generating process here is a function of exogenous covariates and all previous observed treatment levels. All simulations use the identical outcome variable model.

Specifically, in the first set of simulations, for time j, we use the covariates $X_{ij} = (Z_{ij1} \cdot U_{ij}, Z_{ij2} \cdot U_{ij}, |Z_{ij3} \cdot U_{ij}|, |Z_{ij4} \cdot U_{ij}|)^\top$ where each Z_{ijk} is an i.i.d. draw from the standard normal distribution, and U_{ij} is constructed as $U_{ij} = 2 + (2T_{i,j-1} - 1)/3$ for $j = 2, 3$ and $U_{ij} = 1$ for $j = 1$. The treatment assignment model is given by $\Pr(T_{ij} = 1) = \text{expit}\{-T_{i,j-1} + \gamma^\top X_{ij} + (-1/2)^j\}$ where $\gamma = (1, -0.5, 0.25, 0.1)^\top$ and $T_{i0} = 0$. Finally, the outcome model is defined as $Y_i = 250 - 10 \cdot \sum_{j=1}^{3} T_{ij} + \sum_{j=1}^{3} \delta^\top X_{ij} + v_i$ where $\delta = (27.4, 13.7, 13.7, 13.7)^\top$ and v_i is a normal disturbance with mean zero and standard deviation five. To consider the functional form misspecification, we use the following non-linear transformation of the covariates, $X^*_{ij} = (X^3_{ij1}, 6 \cdot X_{ij2}, \log(X_{ij3}), 1/X_{ij4})^\top$ and estimate the treatment assignment model with these covariates. The misspecification was selected to induce skew in the transformed covariates. In our experience, logistic regression estimated propensity scores can be particularly sensitive to misspecifications with skewed covariates.

In the second set of simulations, we consider a misspecification of lag structure. The current treatment level is generated from a function of all previous observed treatment levels and covariates, but only the covariates from the current period and the treatment from the most immediately previous time period are used in estimating the weights. As with the first two simulations, we also consider the misspecification of functional forms using a nonlinear transformation. Specifically, the treatment assignment in the second set of simulations is given by $\Pr(T_{ij} = 1) = \text{expit}\{\sum_{j'=1}^{j} (T_{i,j'-1} + \gamma^\top X_{ij'})/2^{j-j'} + (-1/2)^j\}$. The true treatment assignment model is a function of the entire covariate and treatment history for each observation, but each method is applied using the most immediate covariates and treatment. In order to generate our covariates for this set of simulations, we adjust U_{ij} such that $U_{ij} = \prod_{j'=1}^{j-1} \{2 + (2T_{ij'} - 1)/3\}$ for $j = 2, 3$ and $U_{ij} = 1$ for $j = 1$. The new set of covariates are then constructed as $X_{ij} = (Z_{ij1} U_{ij}, Z_{ij2} U_{ij}, |Z_{ij3} U_{ij}|, |Z_{ij4} U_{ij}|)^\top$

so that they are a function of all past treatments. The outcome model is the same as the one used for the first set of simulations except that the definition of X_{ij} is different. As before, we assess each methods' performance when using the correct covariates, X_{ij}, and the covariates after a mild nonlinear transformation, $X^*_{ij} = \{(Z_{ij1}U_{ij})^3, 6 \cdot Z_{ij2}U_{ij}, \log|Z_{ij3}U_{ij}|, 1/|Z_{ij4}U_{ij}|\}^\top$.

To evaluate the performance of our proposed CBPS methodology, we simulate 2,500 data sets using the aforementioned data generating processes. We then fit a logistic regression model (GLM) as the treatment assignment model independently for each time period using correct and incorrect model specifications as discussed above. We also fit the same exact logistic model using the proposed CBPS estimation but in two ways: first with the fully efficient covariance matrix (CBPS) and with its low-rank approximation (CBPS-Approximate). Finally, the marginal structural model (MSM) weights are constructed from each of the fitted models and then we regress the outcome variable on all three treatment variables using the stabilized MSM weights. The resulting regression coefficients are then compared with the numerical estimates of true regression coefficients obtained from a large number of simulations with the true treatment assignment probabilities.

Figure 2 presents the results from the first set of simulations where the misspecification of treatment assignment model is confined to the functional form and the correct lag structure is maintained. The first three columns show that the bias (upper two rows) and root mean squared error or RMSE (bottom two rows) for the estimated regression coefficients of the three treatment variables (one for each of the three time periods, i.e., $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$, respectively) from the MSM. That is, we use a weighted linear regression where the outcome is regressed on three treatments using the MSM weights. The final column presents the bias and RMSE for the estimated mean potential outcome, $\mathbb{E}(Y_i(t_1, t_2, t_3))$, averaged across eight unique treatment sequences.

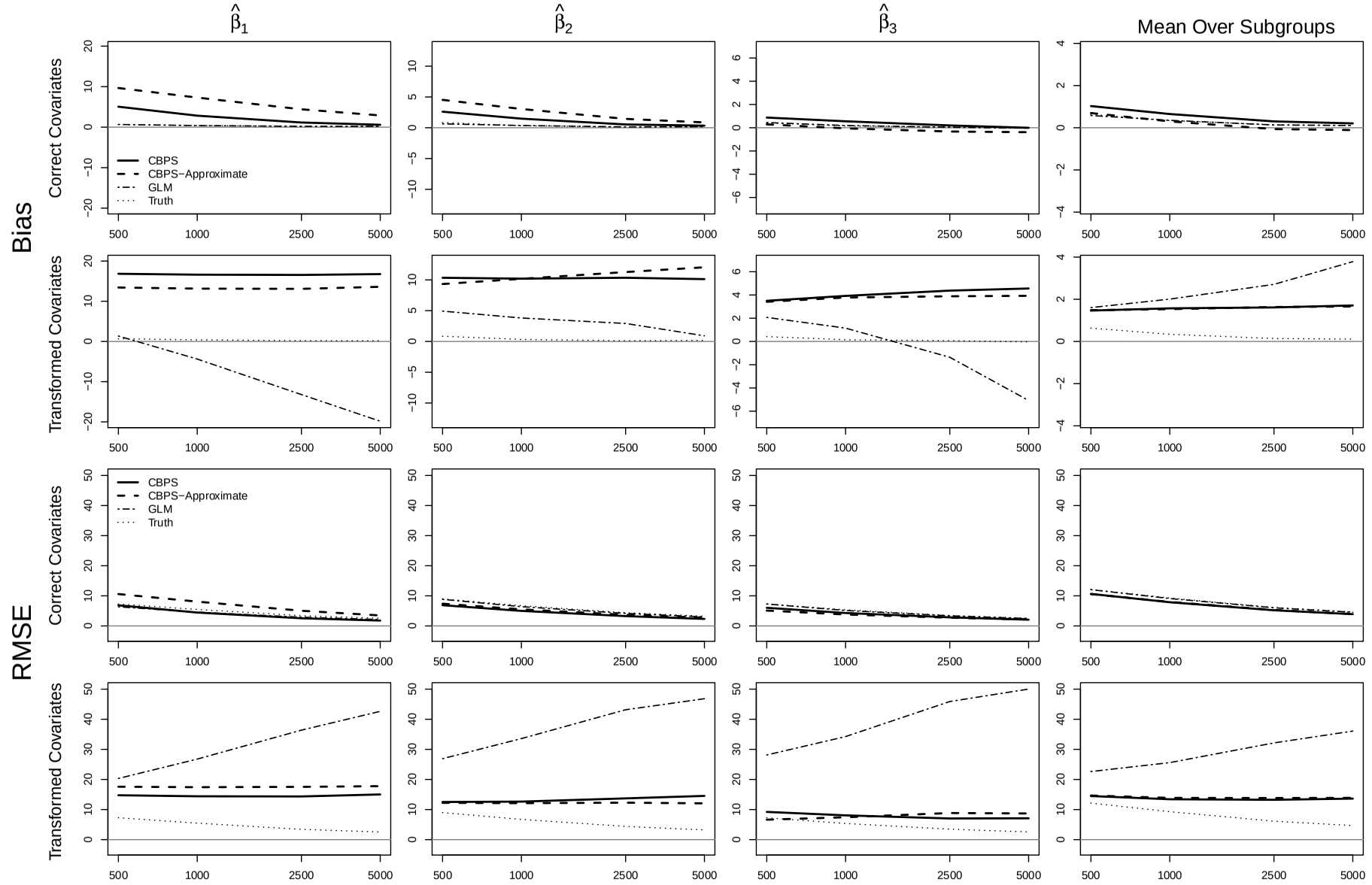


Figure 2: Impact of Treatment Assignment Model Misspecification based on Simulations with Correct Lag Structure. Two cases are considered where the treatment assignment model is either correctly specified (first and third rows) or misspecified (second and fourth rows). In the latter scenario, the regression model is misspecified while the correct lag structure is maintained. The first three columns show that the bias and root mean squared error (RMSE) for the estimated regression coefficients of the three treatment variables (one for each of the three time periods) from the marginal structural model. The final column presents the bias and RMSE for the estimated mean potential outcome, $\mathbb{E}(Y_i(t_1, t_2, t_3))$, averaged across eight unique treatment sequences. Overall, CBPS (thick solid lines) and CBPS with low-rank approximation (thick dash lines) outperform the GLM (thin dot-dash lines) when the model is misspecified. The dotted lines represent the results for the estimates based on the true weights.

These estimates are obtained by calculating the weighted average of the outcome using the subset of data for each treatment sequence.

When the treatment assignment model is correctly specified, all methods have small bias (the first row) and small RMSE (third row) for all quantities of interest. For one parameter in a small sample size, CBPS with the low-rank approximation (thick dash line) has a greater bias than other methods. It is also interesting to note that CBPS is slightly more efficient than the GLM. This finding is consistent with the theoretical result of Hirano *et al.* (2003), which implies that overparameterizing the propensity score model by adding moment conditions can sometimes lead to efficiency gains. However, when the model is misspecified, the bias and RMSE are large and even grow in sample size for GLM (thin dot-dash lines). In contrast, CBPS with the fully efficient covariance matrix (thick solid line) and CBPS with the low-rank approximation have much smaller bias and RMSE across parameters. Unlike the GLM, both the bias and RMSE do not grow in sample size, thereby outperforming the standard estimation technique.

In the first row of Figure 3, the misspecified lag structure induces noticeable bias across all methods, with the CBPS methods showing modest gains in bias (first row) and RMSE (third row). When the lag structure is misspecified and additionally the covariates are transformed (second and fourth rows), the standard GLM estimation leads to much larger bias and RMSE and this bias increases in the sample size. In contrast, the CBPS methods minimize the impact of model misspecification and stays within a reasonable range for bias and RMSE across all quantities of interest.

5 An Empirical Application

We now illustrate the proposed methodology through an empirical application. Blackwell (2013) has applied the MSM to the data from political science in order to estimate

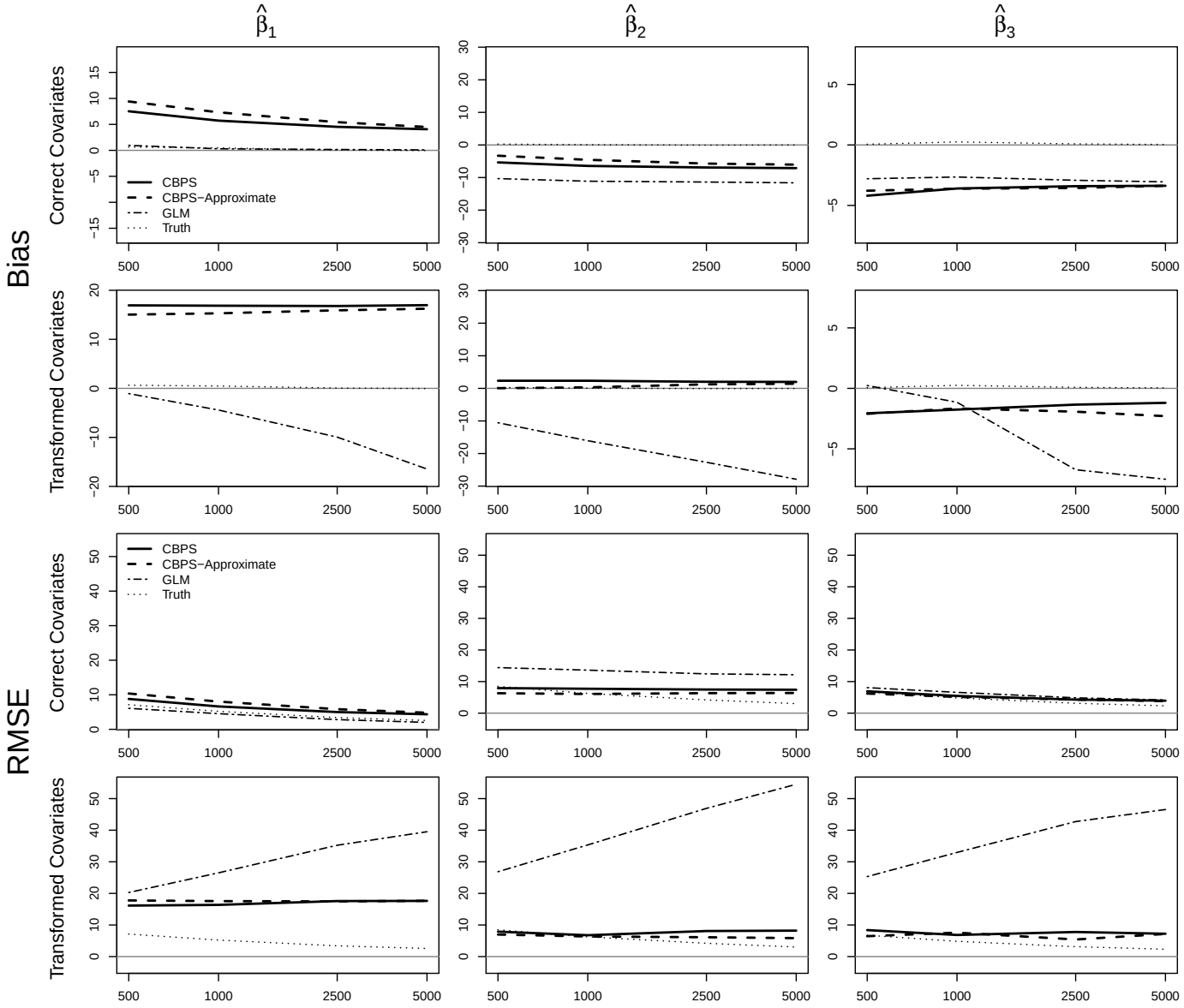


Figure 3: Impact of Treatment Assignment Model Misspecification based on Simulations with Incorrect Lag Structure. Two cases are considered. In the first scenario, the lag structure is incorrectly specified. In the latter scenario, additionally the functional form is misspecified by transforming covariates. The first three columns show that the bias and root mean squared error (RMSE) for the estimated regression coefficients of the three treatment variables (one for each of the three time periods) from the marginal structural model. The final column presents the bias and RMSE for the estimated mean potential outcome, $\mathbb{E}(Y_i(t_1, t_2, t_3))$, averaged across eight unique treatment sequences. Overall, CBPS (thick solid lines) and CBPS-Approximate (thick dash lines) outperform the GLM (thin dot-dash lines) when the model is misspecified. The dotted lines represent the results for the estimates based on the true weights.

the impact of negative advertisements on election outcomes. Here, we analyze a subset of his data. Specifically, we examine the five weeks leading to the elections, using a total of 58 U.S. Senate and 56 gubernatorial candidates from 114 races that were held during the years 2000, 2002, 2004, and 2006. During this time, there were 126 races total in which ads were aired during the last five weeks. Five races are dropped due to missing data, and, following the original author, seven additional non-competitive races were dropped so as to make the common support assumption more credible.

In each week $t = 1, \ldots, 5$ leading up to the election, candidate i may run negative campaign ($T_{it} = 1$) or remain positive ($T_{it} = 0$). The time-varying covariates X_{it} include the Democratic share of the polls, proportion of voters undecided, campaign length, and the lagged and twice lagged treatment variables for each week. In addition, we use the time-invariant covariates including baseline Democratic voteshare, baseline proportion undecided, and indicators for election year, incumbency status, and type of office. The original study fit a single logistic regression model to all time periods, including a linear time trend. In contrast, we allow the coefficients in the model to be different for each time period. We find that the added flexibility yields significantly better covariate balance.

We consider three approaches: the logistic regression, the CBPS based on the optimal covariance matrix (CBPS), and the CBPS based on the low-rank approximation (CBPS-Approximate). For the CBPS and CBPS-approx approaches, we use the two-step and continuously updating GMM estimators, respectively. Note that the computational time for the CBPS was seven times as long as that for the CBPS-Approximate.

We begin by assessing the degree of covariate balance achieved by the logistic regression and the CBPS. Since there are twelve covariates per time period, we have a total of $1548 = \sum_{j=1}^{5} 12 \times (2^5 - 2^{j-1})$ different covariate balancing conditions. As

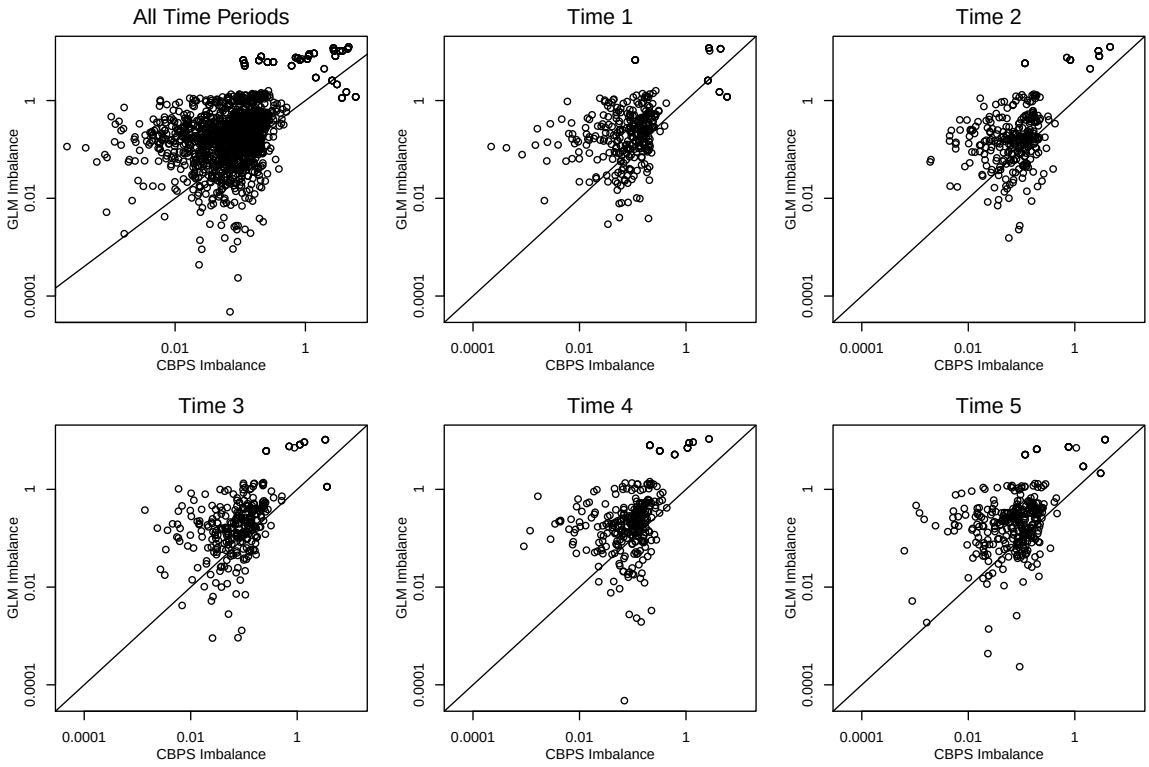


Figure 4: Absolute Imbalance for Each Covariate Balancing Condition by Time Period. The results are compared between the logistic regression (vertical axis) and the CBPS (horizontal axis). The imbalance for all balance conditions appear together in the top left plot, and they are broken out by time period in the remaining five plots. Points above the 45° line indicate that a better balance is achieved for the CBPS than the logistic regression. Overall, the covariate balance achieved by the logistic regression tends to be worse than the CBPS.

shown in Section 3, these moment conditions are implied by the fact that at each time period, conditional on the past treatment history, the MSM weights should balance covariates across all future potential treatment sequences. We characterize the imbalance as the absolute value of the balance conditions for each balance condition $\mathbf{G}$.

Figure 4 presents the absolute imbalance for each covariate balancing condition based on the logistic regression (vertical axis) and the CBPS (horizontal axis). The imbalance of all balance conditions appear together in the top left plot, and they are broken out by week in the remaining five plots. Points above the 45° line indicate that a better covariate balance is achieved for the CBPS than the logistic regression.

	GLM	CBPS	CBPS (Approximate)	GLM	CBPS	CBPS (Approximate)
(Intercept)	55.69*	57.14*	57.38*	55.61*	57.08*	57.31*
	(4.58)	(1.83)	(2.24)	(3.06)	(1.67)	(1.96)
Negative (time 1)	3.02	5.86	2.80			
	(4.51)	(5.27)	(4.72)			
Negative (time 2)	3.54	2.71	4.81			
	(9.61)	(9.21)	(9.80)			
Negative (time 3)	−2.82	−3.93	−4.45			
	(12.44)	(10.89)	(13.62)			
Negative (time 4)	−8.22	−9.72	−8.69			
	(10.19)	(7.75)	(10.81)			
Negative (time 5)	−1.62	−1.98*	−1.91			
	(0.96)	(0.95)	(0.99)			
Negative (cumulative)				−1.18	−1.35*	−1.43*
				(0.67)	(0.38)	(0.45)
R^2	0.05	0.14	0.10	0.03	0.10	0.08
F statistics	1.03	3.44	2.49	3.08	12.43	10.06

Table 3: Estimated Average Causal Effects of Negative Advertising on Candidate's Voteshare. The left three columns present the estimated average causal effects of the time-specific decision to engage in negative advertising. The right three columns contain the estimated causal effects of the cumulative number of periods that the candidate has gone negative. All results are based on weighted linear regressions. The weights are estimated using the logistic regression (GLM), the CBPS with the optimal covariance matrix, and the CBPS with low-rank approximation (CBPS-Approximate). $*$ indicates statistical significance at the 0.05 level.

The logistic regression produces greater imbalance more than 78% of the time, and this pattern is consistent over time, ranging from 75.6% in time 2 to 81.3% in time 4. Relative to the CBPS, the logistic regression has both a greater average absolute imbalance (0.84 versus 0.24) and a larger spread in absolute imbalance (2.05 versus 0.65).

Table 3 presents the estimated impact of negative advertising on candidate's voteshare. The left three columns present the estimated average causal effects of the time-specific decision to engage in negative advertising. The right three columns contain the estimated causal effects of the cumulative number of periods that the candidate has gone negative. All results are based on weighted linear regressions. The weights are estimated using the logistic regression (GLM), the CBPS with the optimal co-

variance matrix, and the CBPS with low-rank approximation (CBPS-Approximate). While these results are somewhat similar across the methods, there are some differences. In particular, when using the CBPS and CBPS-Approximate, the effects of negative advertisement are estimated appear to be more strongly negative than the GLM: the magnitude of estimates effects is larger and standard errors tend to be smaller. The R^2 and F statistics are also greater when the weights are estimated using the CBPS and CBPS-Approximate. Finally, the low-rank approximation approach for the CBPS does not alter the results significantly.

6 Concluding Remarks

In this paper, we have extended the CBPS methodology of Imai and Ratkovic (2014) to the estimation of inverse probability weights for marginal structural models (MSMs), a popular tool in the analysis of longitudinal data. The proposed methodology estimates these weights by improving the resulting covariate balance. This is an important advantage because checking covariate balance, after fitting the treatment assignment models, is a difficult task even when the number of time periods is moderate. As a result, detecting model misspecification is much more challenging in longitudinal data settings than simple cross-section data settings.

In addition, because the MSM weights are constructed by multiplying the inverse of the estimated propensity scores from each time period, MSMs can be highly sensitive to the misspecification of treatment assignment models. In contrast, the CBPS methodology provides a robust estimation method for inverse probability weights by maximizing covariate balance. Our simulation and empirical studies illustrate the effectiveness of the proposed method over the standard maximum likelihood estimation.

One possible future research agenda is the extension of the proposed methodology

to non-parametric estimation using empirical likelihood. The MSM weights are often estimated using a parametric model but better covariate balance might be achieved by using a more flexible estimation approach. Another important and yet unresolved question concerns the selection of covariate balancing conditions when there exist many such conditions. As we have shown, the number of covariate balancing conditions grow exponentially as the number of time periods increases. Under this scenario, the data will become sparse and some treatment sequences have extremely small number of observations. Here, the application of moment selection methods may be useful. We plan to investigate how the proposed CBPS methodology performs in such a situation and develop effective strategies for addressing this issue.

References

Blackwell, M. (2013). A framework for dynamic causal inference in political science. *American Journal of Political Science* **57**, 2, 504–520.

Box, G. E., Hunter, J. S., and Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley-Interscience, New York, 2nd edn.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* **168**, 6, 656–664.

Fong, C., Ratkovic, M., and Imai, K. (2014). CBPS: R package for covariate balancing propensity score. available at the Comprehensive R Archive Network (CRAN). `http://CRAN.R-project.org/package=CBPS`.

Good, I. J. (1958). The interaction algorithm and practical Fourier analysis. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **20**, 2, 361–372.

Graham, B. S., Campos de Xavier Pinto, C., and Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies* **79**, 3, 1053–1079.

Hainmueller, J. (2012). Entropy balancing for causal effects: Multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* **20**, 1, 25–46.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 4, 1029–1054.

Hirano, K., Imbens, G., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 4, 1307–1338.

Howe, C. J., Cole, S. R., Chmiel, J. S., and Muñoz, A. (2011). Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. *American Journal of Epidemiology* **173**, 5, 569–577.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **76**, 1, 243–263.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussions). *Statistical Science* **22**, 4, 523–539.

Lefebvre, G., Delaney, J. A. C., and Platt, R. W. (2008). Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine* **27**, 18, 3629–3642.

Mortimer, K. M., Neugebauer, R., van der Laan, M., and Tager, I. B. (2005). An application of model-fitting procedures for marginal structural models. *American Journal of Epidemiology* **162**, 4, 382–388.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). *Statistical Science* **5**, 465–480.

Robins, J. (1999). *Statistical Models in Epidemiology, the Environment and Clinical Trials (eds. M. E. Halloran and D. A. Berry)*, chap. Marginal Structural Models Versus Structural Nested Models as Tools for Causal Inference, 95–134. Springer, New York.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods: Application to control of the healthy worker survivor effect. *Mathematical Modeling* **7**, 1393–1512.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 5, 550–560.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29**, 159–183.

Yu, Z. and van der Laan, M. (2006). Double robust estimation in longitudinal marginal structural models. *Journal of Statistical Planning and Inference* **136**, 3, 1061–1089.